INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

Vol 7, Issue 11, (2017) E-ISSN: 2222-6990

Exploring the Conceptual and Psychometric Properties of Alternative Assessment using the Rasch Measurement Model

Nor Hasnida Che Md Ghazali

Faculty of Human Development, Universiti Pendidikan Sultan Idris, Malaysia

DOI Link: http://dx.doi.org/10.6007/IJARBSS/v7-i11/3559

Published Date: 21 November 2017

Abstract

The aim of this study is to develop and provide the evidence of psychometric evaluation of an instrument to measure the implementation of an alternative assessment in the context of Islamic education teachings at primary school level. In order to conceptualize alternative assessment implementation, a 32-item questionnaire was designed and administered to a group of 105 primary school teachers teaching Islamic education subject. The respondents were selected through purposive sampling. Content validity is conducted by two experts in the field from an education university. Then, the construct validity and reliability are assessed using the Rasch Measurement Model by identifying the scale rating, uni-dimensionality, item polarity, item fit, item difficulty index, item reliability, person reliability and separation index. The findings revealed that the structure calibration value with the difference of threshold is between 1.4 and 5.0. The values of PMC range from 0.46 to 0.72. All items are in the range of 0.6 to 1.4. Item C17 is removed, and it is the most difficult item. Item reliability is 0.88, person reliability is 0.97 and the separation index is 8.82 for items and 3.68 for person. The overall item quality is good. Obviously, this study does look into the implementation of AA from the sources available, the teaching strategies used by teachers and also the challenges they face in the Malaysian educational context. This validated instrument could be used for real study and also as a self-assessment tools for teachers. Teachers could then determine their strength and weaknesses.

Keywords: Construct Validity; Construct Reliability; Alternative Assessment; The Rasch Model.

Introduction

A very high quality of an educational system should be the main backbone of a country's development. An educational process has to be implemented with full of responsibility, efficient and focused so that it would give a positive impact to the knowledge development, skills and attitude of the students (Rohaya & Mohd Najib, 2006). In education, there are three main elements involved; teaching pedagogy, curriculum and assessment. All of the three are interrelated with each other. All of them have to be concordance with the international benchmark to ensure that the students acquire knowledge and skills which suits the 21st century and also could get along with the spirit of long-life learning.

In addition, the five-year plan in the Eleventh Malaysian Plan (2016-2020) is the final leg before we enter the arena of developed nations. Malaysia is targeting to achieve The Gross National Income per capita of 15 thousand US dollars, which is the level of income of an advanced nation by international standards (PMO, 2015). However, the definition of an advanced nation should not be solely based on per capita income. There are Six Strategic Thrusts altogether in the 11th Malaysian Plan which include Inclusivity, Wellbeing of the Rakyat, Human Capital, Green Growth, Infrastructure and Innovation and Productivity. With regard to the strategic thrust on Inclusivity especially for youth, who are the nation's hope and an important asset, the Government realizes their potential through capacity building, education, skills training, entrepreneurship, sports and volunteerism by implementing the new National Youth Policy. This seems to be consistent with the interest of this study. Similarly, education systems around the world are also going through reform concerning students' performance (Fullan, 2011). In order to improve the teaching and learning process, it is important to consider the three main aspects in education which are the curriculum, instruction and assessment (Young and Giebelhaus, 2005). Assessment, being the focus of this study will be looked into in detail. In order to improve the assessment system in guaranteeing students' learning outcome, Malaysia has come out with the latest Plan which is called the Pelan Pembangunan Pendidikan Malaysia (2013-2025). In 2011 the government has implemented School-based Assessment system which introduces formative assessment to our assessment system. This new assessment system which includes alternative assessment (AA) strategies is a new format which is more holistic and strong. It is also concordance with a new standard curriculum. The question is, with this new form of AA, are teachers ready to implement it in their teachings? Study has revealed that most teachers agree and support the AA approaches, but they are worry on how to plan and implement AA and also how to evaluate students' learning outcome (Salmiah, 2013). Furthermore, teachers have to have certain knowledge and skills in implementing AA because the failure in mastering them could influence the degree in the confidence level of teachers (Sasmaz, 2014).

Assessment is very important in classroom and learning (Stiggins, 2002). More than that, assessment plays a critical role for education policy makers and practitioners involved with accountability of students learning and instruction of teachers (Danielson, 2001). A good assessment process should be integrated into the teaching and learning process (Russel & Airasian, 2012) and not be separated from it. Assessment and evaluation of students' performance seem to be one of the most challenging task for teachers (Maslovaty & Kuzi, 2002) All these while, teachers are using traditional assessments. Traditional assessments are the form of assessments whereby teacher sets and defines the task and determines how performance should be evaluated (Gibbs & Simpson, 2004) which include standardized and classroom achievement tests with mostly closed-ended item, such as true or false, multiple choice and fill-in-the blanks (Belle 1999). However, few researchers such as Black (1998), Broadfoot (1996), Lambert and Lines (2000) and Shepard (2000) found that traditional assessment could not measure students' achievement in an effective way (Wikstrom, 2007). In the early 1990s, many researchers started to be concerned about the alternative form of assessments which presents a process that provides an opportunity for a meaningful integration of curriculum, instruction and assessment (Wikstrom, 2007).

By definition, AA is any type of assessment in which student creates, making or generating a response to a question or task and most probably students are required to produce (Wikstrom, 2007). This is different from traditional assessment in which students merely select a response from a given list of responses. Some researchers consider AA as performance-based assessment or authentic assessment whereby progress of students are measured based on the way the student completes a specified task (Wikstrom, 2007). For this study context, all terms (AA or performance-based or authentic assessment) are used interchangeably as these assessments have one thing in common whereby, each student should generate a response rather than choosing the answer from those given. Two central features of AA are that it is seen as an alternative to the traditional multiple-choice test or standardized achievement test and it involves a direct examination of students' performance on significant tasks that are relevant to life outside of school (Burke, 2005). In short, AA is a form of assessment which seems to fill in the gap that the traditional assessment could not fulfill (Aydin, 2005).

AA strategies allow learners to demonstrate outcomes in different ways like drawing or writing, observing and communicating (Fensham, 2004). AA strategies include open-ended questions, exhibits, demonstrations, hands-on execution of experiments, computer simulations, projects and portfolios (Dietel et al., 1991). As for example, portfolios are a purposeful collection of student work in one or more areas which include student participation in selecting contents, the criteria for judging merit and evidence of student self-reflection (Bailey, 1998). Portfolios require a lot of input and responsibility from the student and a great deal of time commitment from teachers. According to Bhasah Abu Bakar (2006), the difference between AA and traditional assessment are that, in AA; i) the assignments are more related to real life situations; ii) the assignments are more complex and less structured to allow authenticity, and students are allowed to think and come out with various solutions; iii) more time is needed to evaluate due to the difficulty in planning, designing and evaluating students' assignments; and iv) more teacher judgment is needed in evaluating the assignments as they are more complex and authentic.

Although educationists support the shift from traditional assessment to AA, its implementation is facing lots of challenges. Two main challenges are lack of knowledge and lack of skills in practicing AA (Noormarina, 2015). Lack of time, too many students in a classroom, less references on methods and too many documentation tasks also pose a problem (Sazmaz et al., 2011). Metin (2013) states two main themes for the challenges faced by teachers are the problem in determining subject and also criteria. They interviewed teachers and found that they could not decide which subject content to turn into assignments in order for them to measure different level of students. Teachers are not able to explain the importance of the tasks in order to motivate students. Teachers are having problems in determining the types of tasks which is related to the curriculum needs. Problems also could happen during assessing whereby more time is needed, could not assess objectively, teachers are not able to give an effective feedback or teachers do not involve actively during the assessment process. Some teachers do not give continuous feedback to ensure a good quality of assessment.

A valid, reliable and practical instrument is needed in conducting research. A questionnaire could only be useful if it produces meaningful and trustworthy data. Or, in other words when it measures what it is supposed to measure and the measurement is stable in measuring certain concepts. An instrument is valid when it is measuring what is supposed to measure (Muijs, 2011). Reliability on the other hand is defined as 'the extent to which test scores are free from measurement error' (Muijs, 2011). It is a measure of stability or internal consistency of an instrument in measuring certain concepts (Jackson, 2003). The justification of testing the psychometric aspect of this instrument is to provide empirical evidence which providing a credible measurement. Saifudin et al. (2010) state that a meaningful measurement of any variable produces standardized instrument which could be replicated. Until now, there is no study which tests the validity and reliability of this instrument in the context of Islamic education teachings amongst primary school teachers using the Rasch measurement model. The Rasch measurement model is a technique measuring latent traits (Azrilah et al., 2013). The uniqueness of this model is that it could determine whether respondents have a clear understanding on variables. It could also measure not only person but also items, which are measured at one column of a same linear scale (Bond & Fox, 2001). The analysis using the Rasch measurement model has the capability to overcome the weakness from the classical testing theory (CTT) especially in determining item difficulty level aspects. CTT assumes the item difficulty is sample-dependent (Fan, 1998). And, CTT is not able to predict sample performance or ability. In Rasch, sample ability and item difficulty are managed on the same logits scale which enable the sample ability to be predicted (Bond & Fox, 2012). Therefore, by providing the psychometric evaluation of this instrument would fulfil the research objective.

Objectives of the Study

Teachers are trained to develop a valid and reliable assessment instruments during teaching and learning process but the implementation process of AA may affect their AA activities conducted in classrooms (Danielson, 2001). However, until now, in the Malaysian context, there is no thorough study regarding the development of instrument on the implementation of AA in classrooms. Some previous studies focused on the impact of AA students' motivational state and self-efficacy (Sasmaz, 2011) or teachers' perception on AA (Johari et al, 2009). There is a research which develop instrument looking at the teacher practises and the challenges faced by the student teachers in implementing AA (Buldur & Tartar, 2011) but not in the Malaysian context. Obviously, this study does look into the implementation of AA from the sources available for teachers, the teaching strategies used by teachers and also the challenges they face in the Malaysian educational context. Hence, the goal of this present study is to develop a psychometrically sound self-report questionnaire on AA implementation. In addition, the item analysis is conducted to gain the item difficulty index for each measurement. A psychometrically sound (valid and reliable) instrument can be very useful for researchers and educators interested in determining the implementation of AA.

Methodology

This study is a quantitative approach study which involves the collection of data using questionnaire. The questionnaire were administered to 105 primary school teachers teaching Islamic Education subject from 14 secondary schools in Malacca in Malaysia. There are 50 male and 55 female teachers. A set of instrument on the AA implementation was adapted from Normarina (2016). The instrument uses a 4-point Likert scale ranging from '*0*' as *'not confidence at all*' to '*3*' as *'very confidence in doing'* depending on the respondents confidence towards their capability to implement AA. The instrument consists of 32 items. It is used to gauge 3 main constructs; i) sources available in AA implementation (5items); ii) The teaching strategies used during AA implementation (17items); and, iii) challenges faced during AA implementation (10items). Data is analysed using Winsteps software based on the Rasch Measurement Model. According to Green and Frantom (2002), Rasch analysis requires a sample of 100 respondents and 20 items for the data to be considered stable, so this study is suitable enough for that.

Findings and Results

The content validity is evaluated by an expert in educational measurement and educational psychology. Few sentence structures are changed to the existing items. Then, the Rasch Measurement Model is used to assess the validity and reliability of this 32-item instrument.

a. Scale rating

Scale calibration is an important factor in measurement system and data validation. Scale validation would determine whether the data is valid to be analysed. The instrument which is not calibrated could produce data which is not suitable to be analysed. Observed average is consistently increasing from -0.89 to 2.03 (Table 1). This shows the consistency in response. The value for structure calibration must be more than 1.4 but not more than 5.0 (Bond & Fox, 2012). If the difference is less than 1.4, the rating has to be collapsed. If the difference is more than 5.0, the rating has to be separated.

Table 1. Structure Calibration of 4-point scale

Category	Mean	Structure Calibration	Measurement
0	-0.89	None	(-3.81)
1	-0.12	-2.84	-1.68
2	1.89	-0.98	1.67
3	2.03	3.66	(4.33)

b. Uni-dimensionality

Uni-dimensionality is critical in determining an instrument which is measuring in one dimension. An instrument which is not exact in measuring what it supposed to measure could give a confusing outcome. Table 2 shows standard residual variance. The raw variance is 56.3% which is considered medium (Fischer, 2007). This value is near to the expected model value which is 55.8%. According to Azrilah et al. (2013), Rasch analysis needs at least 40 % of raw variance explained by measurement as an indicator of a good uni-dimensionality. So, it has fulfilled the 40% Rasch needs for the instrument needs. In addition, unexplained raw variance in contrast 1 shows 5.3% meaning that it is good and still far away from a standard value which is 15%.

Table 2. Standard residual variance (in Eigenvalue)

	Empirical	(%)	Model (%)
Number of raw variance in observation	73.8	100.0	100.0
Raw variance explained by measurement	41.8	56.3	55.8
Raw variance explained by items	41.8	56.3	55.8
Raw variance explained by respondents	480.6	56.3	55.8

Number of unexplained raw variance	25.0	41.3	42.8
Unexplained raw variance in contrast 1	3.8	5.3	13.1

c. Item polarity
A statistical item showing the correlation results between one point (a response choice) with a continuous variable (scores for all respondents) is called Point Measure Correlation (PMC). In Rasch statistics, the mean square value of the residual item which is sensitive to the items which have failed to relate to the test scores and point-biserial items with very large values are considered. It means that the correlation point size in Rasch is sensitive to the interaction of items (Wright and Stone, 1979). The acceptable critical PMC of an item is 0.2 or more (Pray and Popovich, 1985). A discrimination index of less than 0.2 is weak and more than 0.4 is good (Masey, 2000). A lowest value of PMC for this study is 0.46 whereby the values range from 0.46 to 0.72. This indicates that items could contribute to the measurement of items. These could discriminate or differentiate each level. This indicates that the item discrimination is very good. No items show a negative PMC. A negative or zero value shows that response from items or respondents are not concordance to the constructs or variables (Linacre, 2012).

d. Item Fit
Item fit is checked using MNSQ infit/oufit values. For polytimus scale, infit and outfit of mean square value has to be in the range of 0.6 to 1.4 (Bond and Fox, 2012). If not, the item has to be removed. A value more than 1.40 logit shows that the items are not homogenous with other items in one measurement scale whereas a value less than 0.60 logit shows that there are an overlapping between the construct and other items. For these types of items, they would be best to be removed. In this study, all 32 items are in the range of MNSQ value as suggested by Bond and Fox (2012) so all of the items are fit to be measured.

e. Standardized Residual Correlations
The measurement on the standardized residual correlations is to determine whether the items overlap or not. If the value of the residual correlation is high for the two items, it shows that the items are overlapped. According to Linacre (2012), if the correlation value is more than 0.70, it shows that only one item has to be maintained and the other has to be removed. From the data, there is one set of item overlapped so one item has to be removed. A residual correlation value for item c17 and c13 is 0.81. So, one item is removed which is item c17.

f) Item difficulty index
Item difficulty can be defined as a state of variable continuum from easy to more difficult and it is measured using logits. The item validity is defined via the assessment of item difficulty whereby all of the items are arranged in a hierarchical position to define each construct. In Rasch model, the mean of an item is normally considered as zero (Bond and Fox, 2001). In this study, logits score for item c17, c13, P23 and P9 are +2.29, +2.02, +0.45 and +0.12 respectively. Since the logits score for item c17 is the highest value, so item c17 is the most difficult task to be implemented as perceived by the respondents.

g) Reliability and Separation index
Table 3 shows a high item reliability which is 0.88. This shows that there are enough items to measure. The item quality is also high which shows that they are able to separate individual with a good separation index whereby Person Separation is 3.68. This shows that individual separation index for 105 teachers is divided into 6 strata of an individual ability. However, the real item separation is 2.21 which shows that the 45 items in the instrument could be divided into 3 item strata group with a good Standard Measurement Error (SE) of 0.09. In conclusion, these values show that items have formed variables which are well spread and the position of the item at the logit scale is having a high reliability value.

Table 3. Item and Person Reliability

			INFIT		OUTFIT	
		MEASURE	MNSQ	ZSTD	MNSQ	ZSTD
Item	Mean	0.00	0.91	-0.20	0.89	-0.30
	S.D	0.90	0.27	1.70	0.14	1.60
	Reliability	0.88				
	Separation index	8.82				
Person	Mean	0.72	0.98	-0.78	0.99	-0.50
	S.D	1.56	0.46	2.00	0.56	2.44
	Reliability	0.97				
	Separation index	3.68				

Discussion
Recently, there is no instrument in Malay language which measures the implementation of AA especially in the context of primary school teachers teaching Islamic education subject. Therefore, this study is conducted to develop and measure the psychometric properties of an adapted instrument used to measure AA implementation. This act of validating the instrument to the context which we are interested at is very important as it will give the researcher confidence with the quality of items to be used in real studies for further research. It might be easier to use the existing instrument in the market, but it might not suit the context of this study or this study aims and objectives. Table 4 shows the summary of the result.

Table 4. Summary of the Result

Objective	Research Questions	Findings
To analyze the validity of the instrument	To what extent does the instrument shows the prove related to the validity on: a. measurement scale? b. uni-dimensionality? c. item polarity? d. item fit? e. item difficulty index?	a. Structure calibration value with the difference of threshold between 1.4 and 5.0. b. Standard residual variance (56.3% is explained by the measurement, 5% unexplained raw variance in contrast 1) c. A lowest value of PMC is 0.46 and all values range from 0.46 to 0.72. d. All items are in the MNSQ range of 0.6 to 1.4 but item C13 and C17 has a residual correlation value more than 0.7 so item C17 is removed.

		e. Item C17 is the most difficult item (+2.29 logits).
To analyze the reliability of the instrument	To what extent is the reliability of the instrument in terms of: a. item reliability? b. person reliability? c. separation index?	a. 0.88 b. 0.97 c. 8.82 (item) and 3.68 (teachers)

Validity of the instrument is checked by looking at the scale calibration, uni-dimensionality, item polarity, item fit and item difficulty index. The Rasch model is seen to be able to manage dichotomous or polytimus data into a consistent form of measurement. Uni-dimensionality assumption is fulfilled. Analysis found that the extracted factor has a strength of four items only so it is not considered as one meaningful construct (Fischer, 2007). For item reliability index, it is acceptable if the value is between 0 and 1 or more than 0.8 (Fox & Jones, 1998). Actually, this value is influenced by number of items, the length of the test or the characteristics of the test (Murphy & Davidshofer, 1998). Using this instrument, the implementation of AA is perceived by the teachers through their responses which consider the percentage of teachers' confidence towards three main constructs. Next, the quality of items are good. The items are able to differentiate individuals with a good separation power. Person separation which equals to 3.68 shows individual separation for 105 teachers could be divided into four individual ability strata. Real item separation is 8.82. This shows that all the 32 items used could be divided into nine groups of item strata at a good Standard Error Measurement which is equals to 0.08. This value shows that items have formed a variable which is well spread and the item position at logits scale does have a high reliability value. Item C17 is removed due to the overlapping and it does not affect the content validity of the instrument. Determining the item logits and the value of standardized residual correlation value is important in detecting the presence of overlapping items. This is to avoid the repetition of items assessed similarly by respondents, and this step is important for obtaining optimum number of items.

Implications and Suggestions for Future Research

The results of the literature review and the interview sessions lead the researcher to review the assessment standards, the advantages of using AA and also the challenges that teachers have to face in implementing AA. Furthermore, with this valid instrument, the researcher would be very confident in using this instrument in real study later. The instrument shows that it is reliable to measure and are able to differentiate the levels between teachers based on good separator value. In order to obtain a better understanding of this adapted instrument, it would be advisable to collect data from primary schools in states other than Malacca. This is to test the validity of the study's model across different school samples and the extent to which these can be generalized. This validated instrument could also be used as a self-assessment tools for teachers. Teachers could then determine their strength and weaknesses. Teachers should be given flexibility in assessing their students in a summative way or in a standardized form because in order to improve education quality, assessments should be integrated in the teaching and learning process (Kelvin, 2007).

In future research, the researcher might use the variables developed in this study to look for the interrelationships between variables. This is important as the interrelationships between variables reflect how effective the system are (Nor Hasnida, 2016). This instrument could also be used as a pre-test and post-test when training is conducted. The training could be formal but Slaalvik (2010) believes that stimuli such as praise, encouragement and demonstration can also increase the level of teacher practices towards the implementation of AA. So, in this case, this instrument could also be used informally. Finally, the implementation process of AA should not be taken lightly as Cohen (1995) believed that it could influence the competency of teachers in assessing and it could produce a bigger impact towards students.

Conclusion

The importance of this study is to provide psychometrical characteristics of the implementation of AA amongst teachers teaching Islamic Education subject in primary school level. The findings demonstrate the instrument has adequate psychometric properties for its validity and reliability value. So, this instrument is fit to be used for real study. However, it is better for the instrument to go through further validity and reliability test with larger sample size during real study. Confirmatory factor analysis could also be used to check for the relationship between the variables. The Rasch output has created a paradigm shift in measuring perception by producing more meaningful data and a quality instrument is produced from the individual item checking. The overall item quality is good. Initially it consists of 32 items and finally 31 items are retained. The findings of this study showed the feasibility of using the questionnaire as it has a high validity and reliability value. The Rasch model output has provided statistical evidence for the instrument for future purposes. Thus, this instrument would benefit the stakeholders in assessing the teachers' commitment in implementing AA in primary schools teaching Islamic education subjects.

Corresponding author

Norazilawati Abdullah,
Faculty of Human Development,
Universiti Pendidikan Sultan Idris,
Malaysia.
Email: nora@fpm.upsi.edu.my

References

Aydin, F. (2005). Teacher's opinions and practises about alternative assessment-evaluation. *National Congress on Education Sciences*, 20(1), 775-779.

Azrilah A. A., Mohd Saifuddin M. & Azami Z. (2013). *Asas Model Pengukuran Rasch*. Bangi: Penerbit UKM.

Bailey, K. M. (2015). *Learning about language assessment: dilemmas, decisions, and directions.* Boston: National Geographic Learning.

Black, P. & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education*, Vol. 5, 7-71.

Black, P. (1998). *Testing, Friend or Foe? The Theory and Practice of Assessment and Testing*. US: Falmer Press.

Bond, T. G. & Fox, C. M. (2012). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (2nd.) Now York: Routledge.

Bond, T. & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Broadfoot, P. M. (1996) Education, Assessment and Society: a sociological analysis. *British Journal of Sociology of Education,* 18(3).

Buldur, S. & Tatar, N. (2011). Development of self-efficacy towards using Alternative Assessment Scale. *Asia Pacific Education Review*, 12, 485-495.

Burke, K. (2005). *How to Assess Authentic learning.* Thousand Oaks, CA: Corwin Press.

Cohen, K. M. (1995). *Achieving positive attitudes toward science through alternative assessments*. USA: Newport University.

Danielson, C. (2001). New Trends in Teacher Evaluation, *Educational Leadership*, 58 (5), 12-15.

Dietel, R. J., Herman, J. L., & Knuth, R. A. (1991). What does research say about assessment? NCREL, Oak Brook. Available online: http://www.ncrel.org/sdrs/areas/stw_esys/ 4assess.htm (Accessed 16 October 2017).

Fensham, P. J. (2004). Beyond Knowledge: Other Outcome Qualities for Science Education. Keynote address delivered at the XIth International Organisation for Science and Technology Education Symposium, July 2004, Lublin, Poland.

Fischer, J. (2007). Rasch Measurement Transaction. *Rasch Measurement Transaction*, 21(1), 1095.

Fox, C.M. & Jones, J. (1998). Use of Rasch Modelling in counselling psychology research. *Journal of Counselling Psychology,* 76(4), 569-582.

Fullan, M. (2011). Choosing the wrong drivers for whole system reform. Available at: http://www.michaelfullan.ca/media/13501655630.pdf (Accessed: 3 December 2013).

Gibbs, G. & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education* , 1, 1–31

Gipps, C. V. (2004). *Beyond Testing: towards a theory of educational*. London: The Falmer Press.

Hancock, C.R. (1994). Alternative assessment and second language study: What and why? Available at: https://gse.gmu.edu/assets/docs/forms/mirs/assessment_brochure.pdf (Accessed 16 October 2017).

Green, K. E. & Frantom, C. G. (2002). Survey development and validation with the Rasch Model. Paper presented at the International Conference on Questionnaire Development, Evaluation and Testing, Charleston, SC, November, 14-17.

Hancock, C. R. (1994). *Alternative Assessment and Second Language Study.* Heinle: US.

Lambert, D. & Lines, D. (2000). *Understanding assessment: purposes, perceptions, practice*. London and New York: RoutledgeFalmer.

Jackson, S. L. (2003). *Research Methods and Statistics, A Critical Thinking Approach.* USA: Thomson Wadsworth.

Kelvin, T. (2007). *The case for qualitative approaches to assessment*. In Kelvin Tan (Ed.) Alternative Assessment in Schools: A Qualitative Approach. Jurong, Singapore: Prentice Hall.

Lambert, D. & David, L. (2000). *Understanding assessment: purposes, perceptions, practice.* London and New York: RoutledgeFalmer.

Linacre, J. M. (2007). *A user's guide to WINDTEPS Rasch-model computer programs*. Chicago, Illinois: MESA Press.

Maslovaty, N. & Kuzi, E. (2002). Promoting motivational goals through alternative or traditional assessment. *Studies in Educational Evaluation*, 28(3), 199-222.

Metin, M. (2013). Teachers' difficulties in preparation and implementation of Performance task. *Educational Services: Theory and Practice,* 13(3), 1664-1674.

Muijs, D. (2011). *Doing Quantitative Research in Education with SPSS*. London: SAGE Publications Ltd.

Murphy, K. R. & Davidshofer, C. O. (1998). *Psychological Testing: Principles and Applications*. University of Michigan: Prentice Hall.

Nor Hasnida, C. M. G. (2016). An evaluation of the Implementation of the SBA System in Malaysia. (Doctoral Dissertation). Available at: https://eprints.soton.ac.uk/381724/

PMO (2015). Eleventh Malaysian plan. 2016-2020. Available at: https://www.pmo.gov.my/dokumenattached/speech/files/RMK11_Speech.pdf

Rohaya, T. & Mohd Najib, A. G. (2006). Pembinaan dan Pengesahan Instrumen bagi mengukur tahap literasi pentaksiran guru sekolah menengah di Malaysia, 109-125.

Russell, M. K. & Airasian, P. W. (2012). *Classroom assessment: Concepts and applications* (7thed.) New York: McGraw-Hill

Saidfudin, M., Azrilah, A. A., Rodzo'an, N. A., Omar, M. Z., Zaharim, A. & Basri, H. (2010). Easier Learning Outcomes Analysis using Rasch Model in Engineering Education Research. EDUCATION'10 Proceedings of the 7th WSEAS International Conference on Engineering Education. 442-447.

Salmiah, J. (2013) 'Acceptance towards SBA among agricultural integrated living skills teachers: challenges in implementing a holistic assessment', *Journal of Technical Education and Training,* 5(1), 44-51.

Sasmaz, O. (2014). The alternative assessment evaluation approaches preferred by pre-service teachers and their self-efficacy towards these approaches, *Education and Science*, 39 (173).

Sazmaz, O. & Evrekli, T. (2011). The Science and Technology Pre-Service Teachers' Self-Efficacy Levels and Opinions about Alternative Assessment and Evaluation Approaches. *Educational Services: Theory and Practice*, 11(3), 1690-1698.

Shepard, L. A. (2000). The role of assessment in a learning culture, *Educational Researcher*, 29(7), 4-14.

Slaalvik, E. (2010). Teacher self-efficacy and teacher burnout: A study of relations, *Teaching and Teacher Education,* 26, 1059-1069.

Stiggins, R. J. (2002) 'Assessment Crisis: The Absence of Assessment For Learning', *Phi Delta Kappan,* 83(10), 758-765.

Wikstrom, N. (2007). Alternative Assessment in Primary Years of IBE, Available at: http://www. diva-portal.org/smash/get/diva2:199424/FULLTEXT01.pdf

Young, S. & Giebelhaus, C. (2005) Formative Assessment and Its Uses for Improving Student Achievement. Education Data Management Solutions, STI. Available at: www.cbohm.com/news/STI/STI_White_Paper.pdf (Accessed Nov 2011).